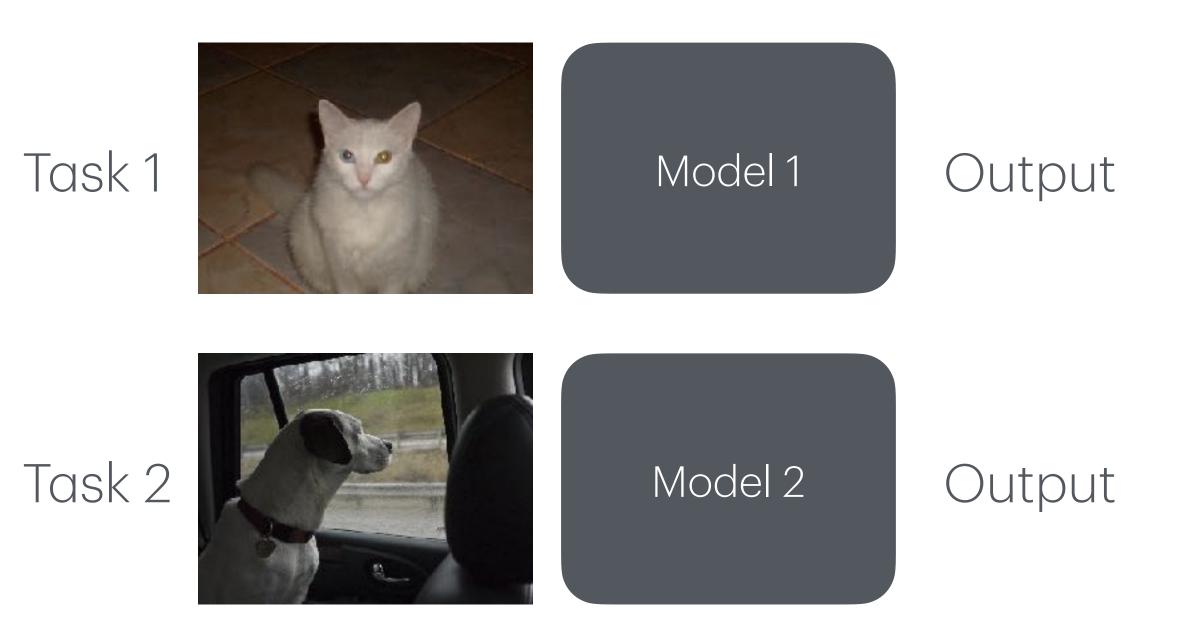
Vision Language Models Image Captioning

Philipp Krähenbühl, UT Austin

Computer vision

"Traditional" Computer Vision 2014 - now



. . .

Unified Computer Vision 2020 - now

Task



Unified Model Output

Image Classification Applications

- Visual search
- Testbed for network design



Apple

Image Captioning

- Similar to classification
 - Richer annotation
 - Cheaper to obtain at scale
 - Alt-text on webpages



An image of an Apple cut into slices.

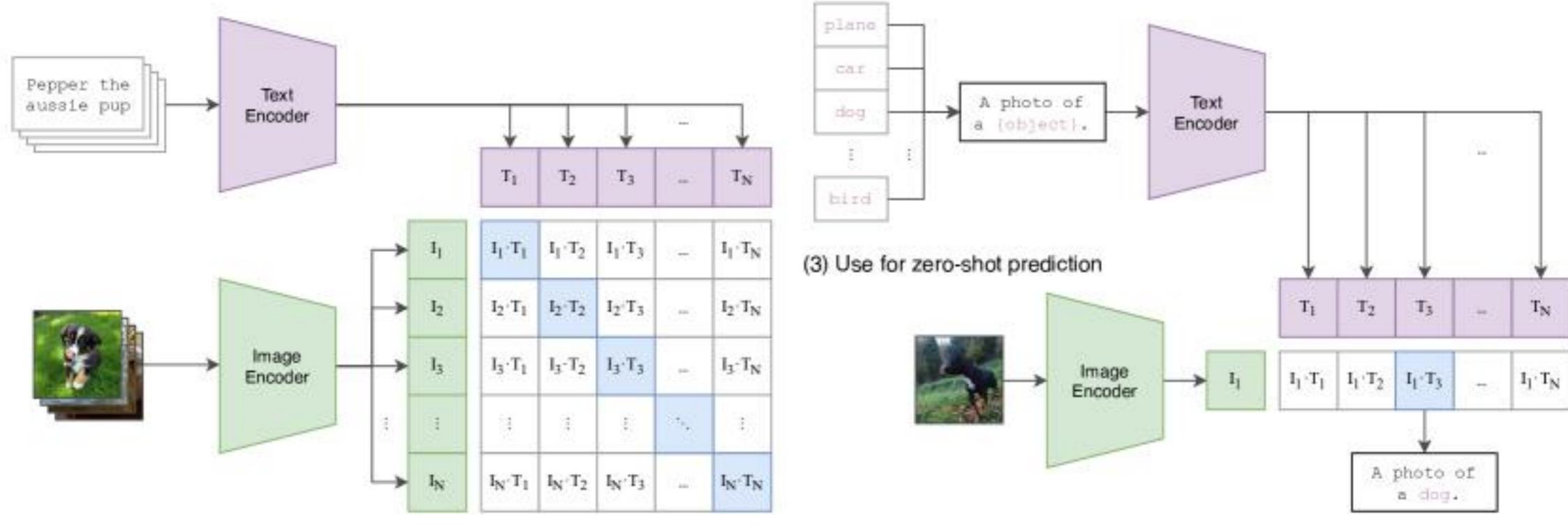
- Download a large curated dataset of image and alt-text
- Learn an image + text embedding
 - With contrastive loss
- Great zero-shot classification performance

Learning Transferable Visual Models From Natural Language Supervision, Radford etal 2021

(*) Contrastive pre-training (2) Create dataset classifier from label text Perper the Text A photo of Text aussis pup Encoder Encoder T₁ T₂ T₃ $\mathbf{i}_{1},\mathbf{i}_{1}=\mathbf{i}_{1},\mathbf{i}_{2}=\mathbf{i}_{1},\mathbf{i}_{3}=\dots=-\mathbf{i}_{1},\mathbf{i}_{N}$ (3) Use for zero-shot prediction T_1 T_2 T_3 _ T_N ▶ I₂ $I_2 T_1 = I_2 T_2 = I_2 T_3$ $\mathbf{I}_2 \cdot \mathbf{T}_N$ 13.T1 12.T2 1.T. $I_1 \cdot T_N$ Image Encoder \rightarrow I₁ I₁·T₁ I₁·T₂ I₁·T₂ $\mathbf{I}_{j}\cdot\mathbf{\Gamma}_{j_{N}}$ A photo of a day. $\mathbf{I}_N \mathbf{T}_1 = \mathbf{I}_N \mathbf{T}_2 = \mathbf{I}_N \mathbf{T}_3$ In. StanfordCars Country211 Food101 Linear Probe C. 70 -Kinetics700 SST2 SUN39 65 -UCF10 ero-Shot BiT-M (ImageNet-21K CLIP. HatefulMerne CIFAR1 CIFAR10 60 55 55 STL FER20 ResNe Caltech10 ImageNe OxfordPets 8 50 · PascalVOC200 Birdsnap JANA 45 MNIST FGVCA rcraft RESISC45 Flowers102 DTD 40 · CLEVRCounts GTSRB PatchCamelyon KITTI Distance 35 EuroSAT 30 · -40 -30 -20 -10 0 10 20 30 40 012 16 - 4 ∆ Score (%) Zero-Shot CLIP vs. Linear Probe on ResNet50 # of labeled training examples per class ImageNet Zero-Shot Dataset Examples ResNet101 CLIP AScore 76.2 76.2 0% ImageNet 70.1 +5.8% 64.3 ImageNetV2 37.7 88.9 +51.2% ImageNet-R Objectiv 32.8 72.3 +39.7% 12-9 16 2 2 ImageNet 25.2 60.2 +35.0% Sketch ImageNet-A 24 77.1 +74.4%



(1) Contrastive pre-training



Learning Transferable Visual Models From Natural Language Supervision, Radford etal 2021

(2) Create dataset classifier from label text

- Good classification model
- Loses many details of image
- Poor localization performance

Learning Transferable Visual Models From Natural Language Supervision, Radford etal 2021

Pepper the Text aussie pup A photo of Text Encoder [object]. Encoder T₁ T₂ T₃ $I_1 \cdot T_1 = I_1 \cdot T_2 = I_1 \cdot T_3$ $I_1 \cdot T_N$ (3) Use for zero-shot prediction $I_2 T_1 = I_2 T_2 = I_2 T_3$ ▶ I₂ $I_2 \cdot T_N$ T₁ T₂ T₃ ... T_N $I_3 \cdot T_1 = I_3 \cdot T_2 = I_3 \cdot T_3$ $I_3 \cdot T_N$ $I_1 \cdot T_1 = I_1 \cdot T_2 = I_1 \cdot T_3$ → I₁ $I_1 \cdot T_N$ Encoder A photo of a dog. IN $I_N^{,\prime}T_1 \quad I_N^{,\prime}T_2 \quad I_N^{,\prime}T_3$ IN TN 75 StanfordCars -28.9 Country211 Food101 Linear Probe C. 23.2 -22.5 70 -Kinetics700 SST2 SUN397 UCF103 65 ero-Shot BiT-M (ImageNet-21K CLIP. HatefulMerne CIFAR1 CIFAR10 00 Score (%) STL FER20. **ResNet** Caltech10 ImageNe OxfordPets g 50 · PascalVOC2007 0.5 Birdsnap -3.2 Ja 45 MNIST -10 FGVCA rcraft RESISC45 Flowers102 DTD 40 · CLEVRCounts GTSRB PatchCamelyon KITTI Distance 35 EuroSAT 30 · -40 -30 -20 -10 0 10 20 30 40 012 16 - 4 8 ∆ Score (%) Zero-Shot CLIP vs. Linear Probe on ResNet50 # of labeled training examples per class ImageNet Zero-Shot ResNet101 CLIP Dataset Examples AScore 76.2 76.2 0% ImageNet 70.1 +5.8% 64.3 ImageNetV2 ImageNet-R 37.7 88.9 +51.2% 32.8 72.3 +39.7% ObjectNe The second 9 % \$ Q ImageNet 25.2 60.2 ±35.0% Sketch

(2) Create dataset classifier from label text

ImageNet-A

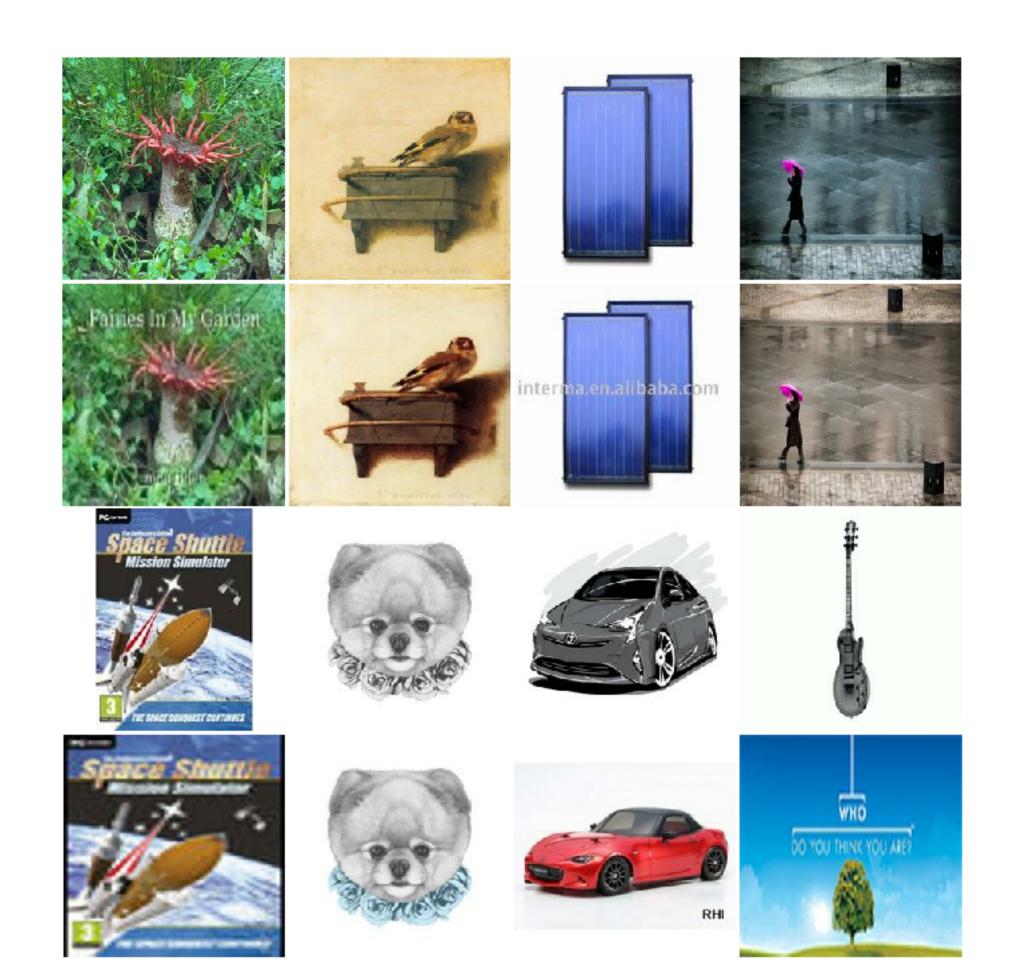
2.7 77.1 +74.4%

(1) Contrastive pre-training

OpenCLIP / LAION

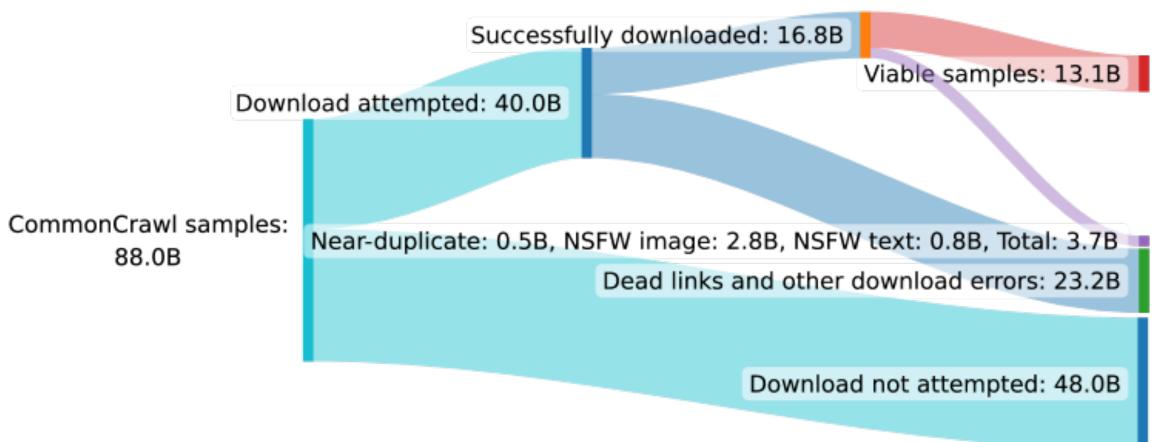
- New image-text dataset: LAION
 - Large scale (up to 2B)
 - DO NOT DOWNLOAD (NSFW)
- Starts to dig into data issues

Reproducible scaling laws for contrastive language-image learning, Cherti etal 2022



DataComp

- Largest image-text dataset yet: 13B images
 - Based on CommonCrawl
- Data filtering as a task
 - Fixed CLIP training
 - Standardized eval



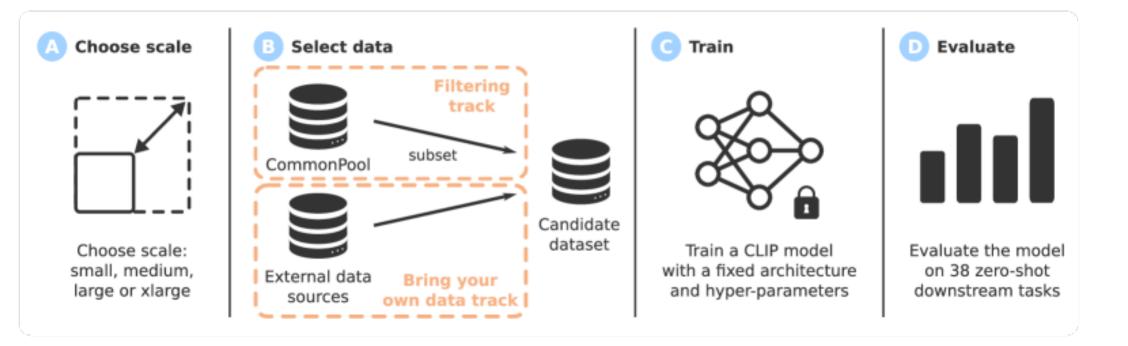


Table 1: Zero-shot performance of CLIP models trained on different datasets. DATACOMP-1B, assembled with a simple filtering procedure on image-text pairs from Common Crawl, leads to a model with higher accuracy than previous results while using the same number of multiply-accumulate operations (MACs) or less during training. See Section 3.5 for details on the evaluation datasets.

Dataset	Dataset size	# samples seen	Architecture	Train compute (MACs)	ImageNet accuracy
OpenAI's WIT [111]	0.4B	13B	ViT-L/14	$1.1 imes 10^{21}$	75.5
LAION-400M [128, 28]	0.4B	13 B	ViT-L/14	$1.1 imes 10^{21}$	72.8
LAION-2B [129, 28]	2.3B	13B	ViT-L/14	$1.1 imes 10^{21}$	73.1
LAION-2B [129, 28]	2.3B	34B	ViT-H/14	$6.5 imes10^{21}$	78.0
LAION-2B [129, 28]	2.3B	34B	ViT-g/14	$9.9 imes10^{21}$	78.5
DATACOMP-1B (ours)	1.4B	13B	ViT-L/14	$1.1 imes10^{21}$	79.2

DataComp

- Very strong baselines
 - CLIP score filtering: discard images with low image-text similarity
 - Text-based filtering: fasttext and caption length filtering
 - Image-based filtering: Cluster clipimage encoder features (and filter according to distance to ImageNet)

DataComp: In search of the next generation of multimodal datasets, Gadre etal 2023

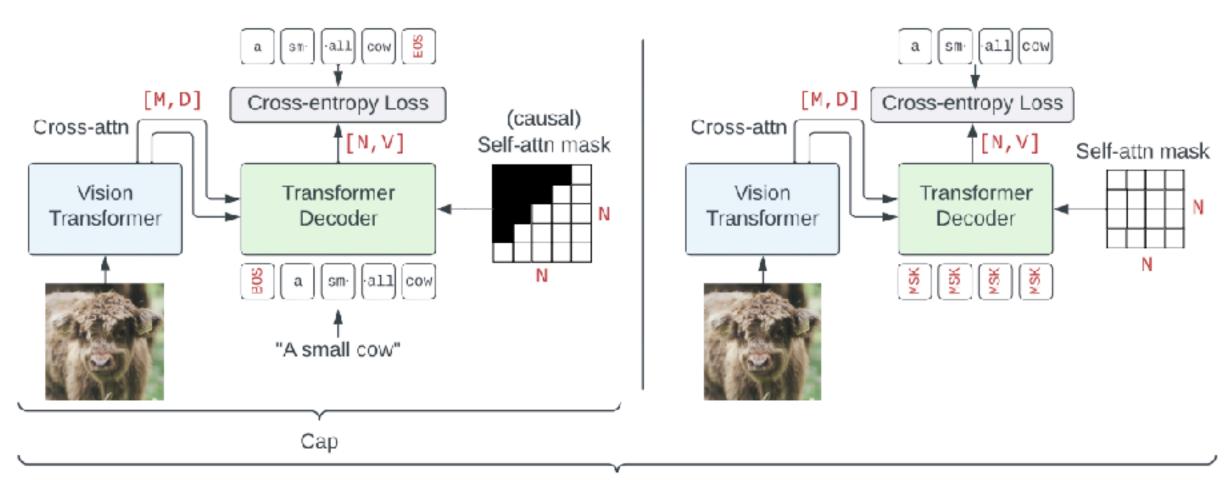
Scale	Filtering strategy	Dataset size	Samples seen	ImageNet	ImageNet dist. shifts	VTAB	Retrieval	Average over 38 datasets
	No filtering	12.8M	12.8M	0.025	0.033	0.145	0.114	0.132
	Basic filtering	3M	12.8M	0.038	0.043	0.150	0.118	0.142
	Text-based	3.2M	12.8M	0.046	0.052	0.169	0.125	0.157
	Image-based	3M	12.8M	0.043	0.047	0.178	0.121	0.159
	LAION-2B filtering	1.3 M	12.8M	0.031	0.040	0.136	0.092	0.133
	CLIP score (L/14 30%)	3.8M	12.8M	0.051	0.055	0.190	0.119	0.173
	Image-based ∩ CLIP score (L/14 30%)	1.4M	12.8M	0.039	0.045	0.162	0.094	0.144
	No filtering	128M	128M	0.176	0.152	0.259	0.219	0.258
	Basic filtering	30M	128M	0.226	0.193	0.284	0.251	0.285
Te	Text-based	31M	128M	0.255	0.215	0.328	0.249	0.307
medium	Image-based	29M	128M	0.268	0.213	0.319	<u>0.256</u>	0.312
	LAION-2B filtering	13M	128M	0.230	0.198	0.307	0.233	0.292
	CLIP score (L/14 30%)	38M	128M	0.273	0.230	0.338	0.251	0.328
	Image-based ∩ CLIP score (L/14 30%)	14M	128M	0.297	0.239	<u>0.346</u>	0.231	0.328
	No filtering	1.28B	1.28B	0.459	0.378	0.426	0.419	0.437
	Basic filtering	298M	1.28B	0.516	0.423	0.446	0.480	0.458
	Text-based	31 7M	1.28B	0.561	0.465	0.465	0.352	0.466
large	Image-based	293M	1.28B	0.572	0.454	0.483	0.479	0.476
	LAION-2B filtering	130M	1.28B	0.553	0.453	0.510	0.495	0.501
	CLIP score (L/14 30%)	384M	1.28B	0.578	0.474	0.538	0.466	0.529
	Image-based ∩ CLIP score (L/14 30%)	140M	1.28B	<u>0.631</u>	<u>0.508</u>	<u>0.546</u>	<u>0.498</u>	<u>0.537</u>
	No filtering	12.8B	12.8B	0.723	0.612	0.611	0.569	0.621
1 1 1 1 1 1	LAION-2B filtering	1.3B	12.8B	0.755	0.637	0.624	0.620	0.636
xlarge	CLIP score (L/14 30%)	3.8B	12.8B	0.764	0.655	0.643	0.588	0.650
	Image-based ∩ CLIP score (L/14 30%)	1.4B	12.8B	<u>0.792</u>	<u>0.679</u>	<u>0.652</u>	0.608	0.663

Training data	Dataset size	# samples seen	ImageNet Acc.	Avg. performa nce (38 datasets)
OpenAl's WIT	0.4B	13B	75.5	0.61
LAION-400M	0.4B	13B	73.1	0.58
LAION-2B	2.3B	13B	73.1	0.59
LAION-2B	2.3B	34B	75.2	0.61
DataComp-1B	1.4B	13B	79.2	0.66

CapPa

- Predict caption from image
 - Auto-regressively (easy for model)In parallel
 - (harder)
- Similar image classification performance than CLIP
- Better captioning and OCR performance
- Poor localization

Image Captioners Are Scalable Vision Learners Too, Tschannen etal 2023

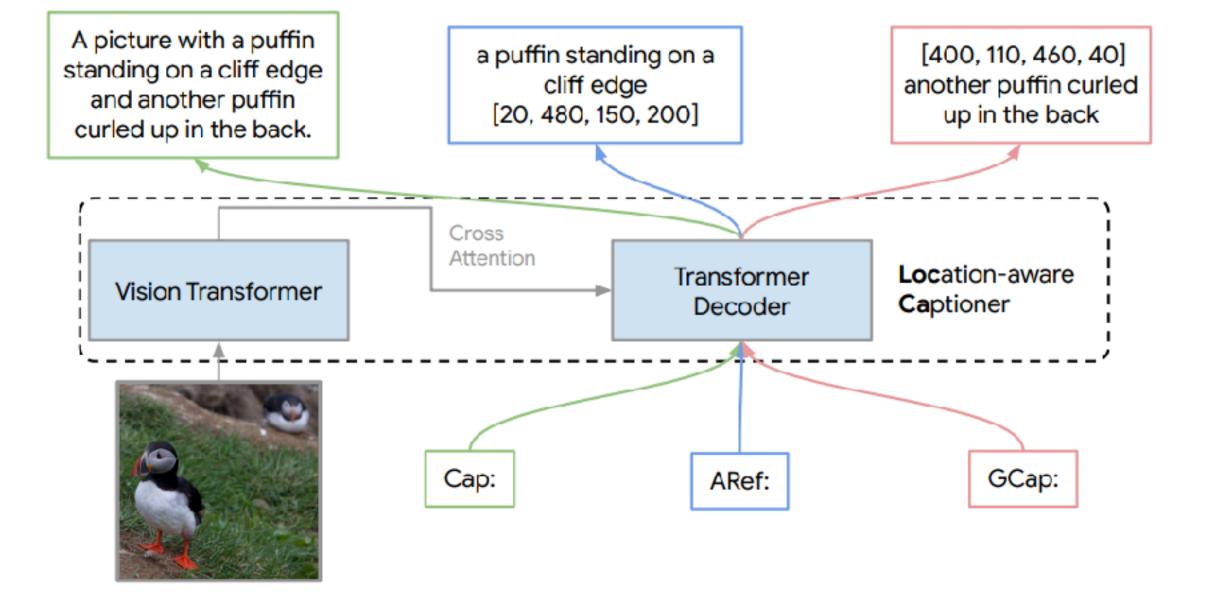


CapPa

B(A) = C(A)

- Add location information to captions
 - Run an off-the-shelf open-vocabulary detector
- Tasks:
 - Captioning
 - Referring expression
 - Grounded captioning
- Similar classification performance
- Better detection, captioning, VQA

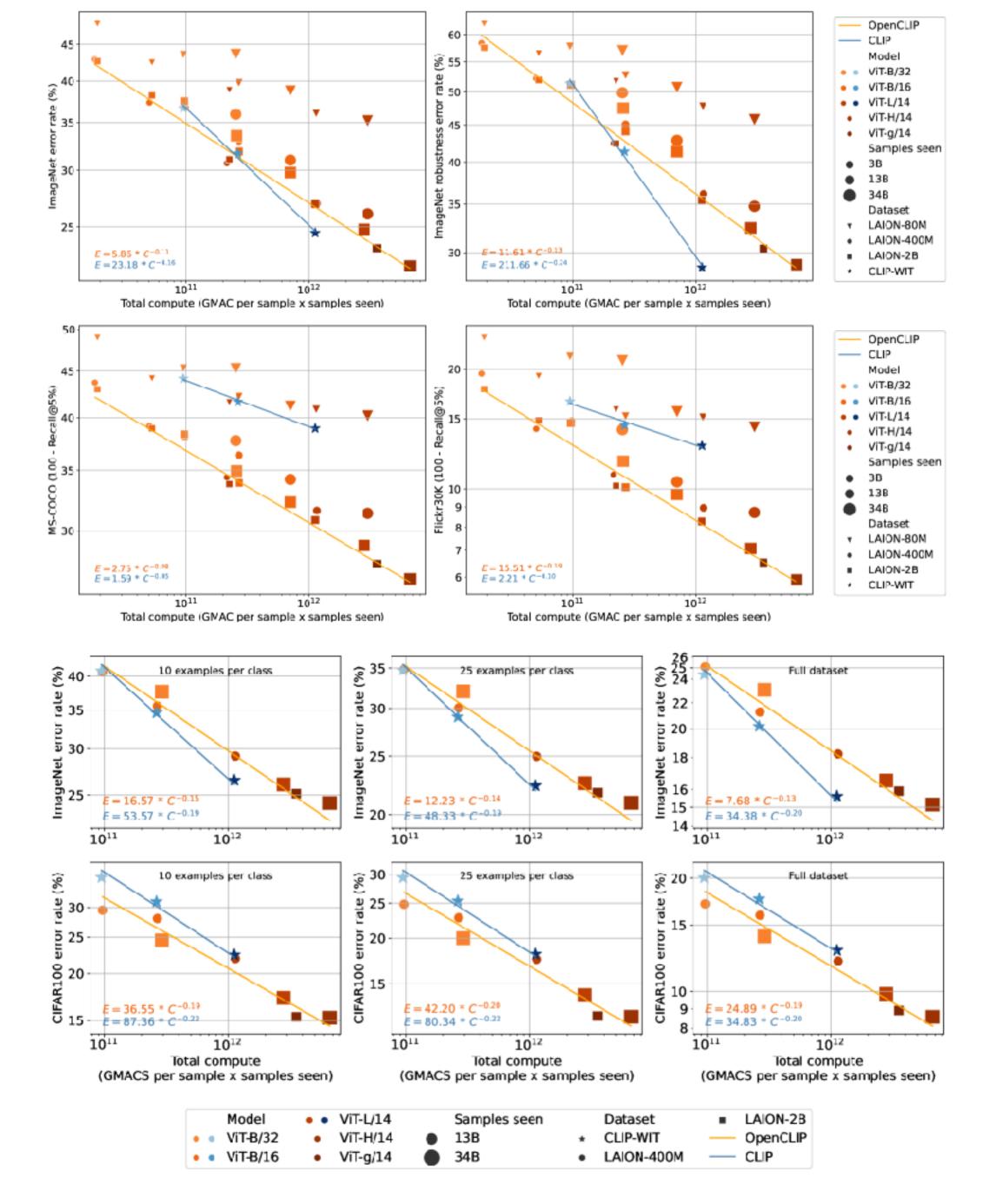
LocCa: Visual Pretraining with Location-aware Captioners, Wan etal 2024



OpenCLIP / LAION

- Open Source replication of CLIP: OpenCLIP
 - Original CLIP may have overfit to ImageNet
- Highlights importance of data

Reproducible scaling laws for contrastive language-image learning, Cherti etal 2022

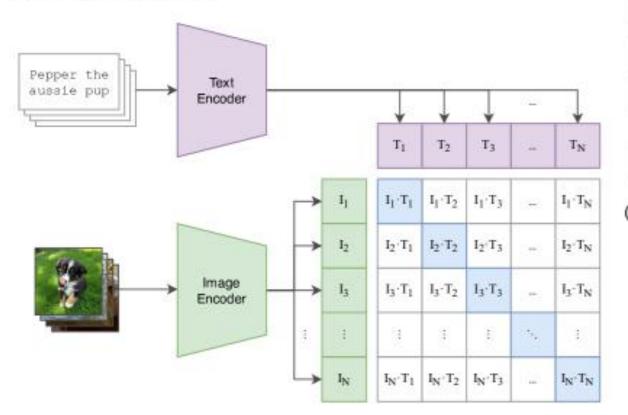


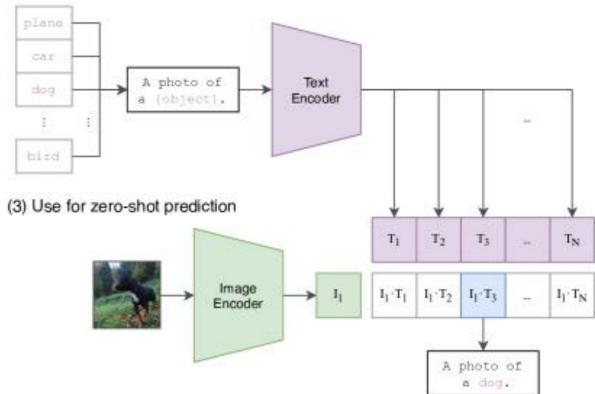
CLIP as a VLM

- Clip maps
 - Images to text
 - Text to images
- Primitive image and text models
- No dialogue

(1) Contrastive pre-training

(2) Create dataset classifier from label text



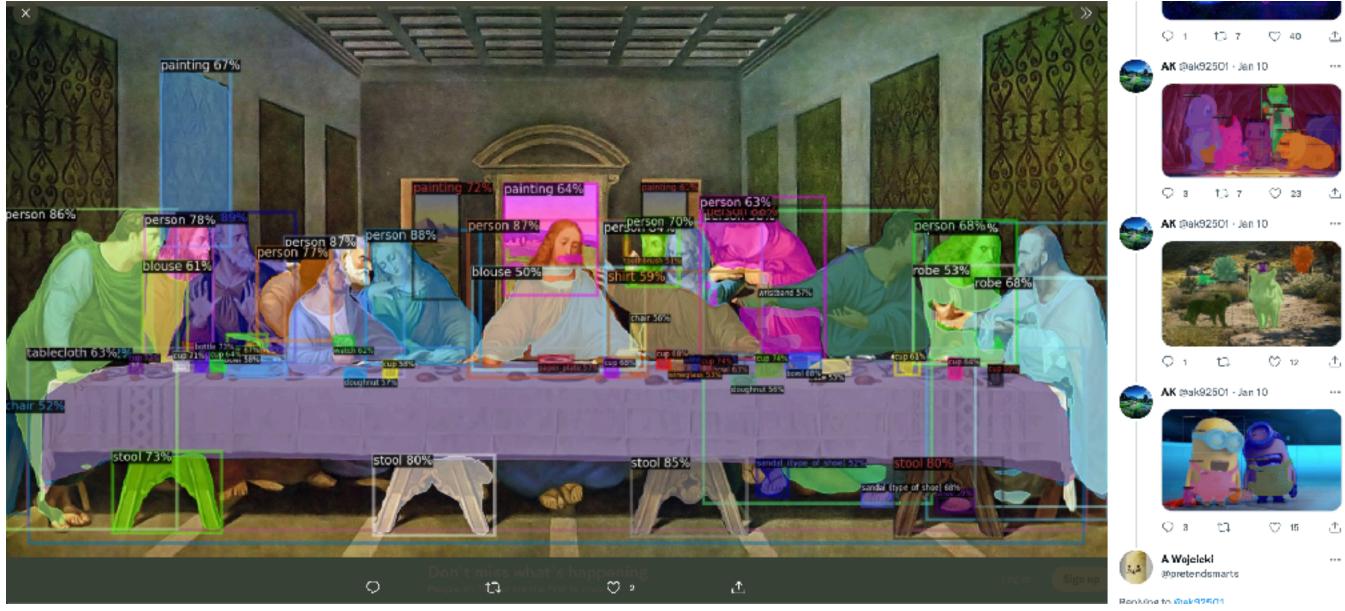


1	

Image Captioning Trend 1

Pre-training on caption datasets

- Much larger (size of internet)
- Richer supervision



Open-Vocabulary Recognition

Trend 2

- Image-encoders, text-encoders, text-encoders, text-decoders are stable
 - Transformer
 - Form of ViT

Open-Vocabulary Recognition

Trend 3

Image data is exhausted, but information is not

- Datasets no longer grow significantly
- What is annotated still grows

References

- Learning Transferable Visual Models From Natural Language Supervision, Radford etal. 2021
- Reproducible scaling laws for contrastive language-image learning, Cherti etal. 2022
- DataComp: In search of the next generation of multimodal datasets, Gadre etal. 2023
- Image Captioners Are Scalable Vision Learners Too, Tschannen etal. 2023
- LocCa: Visual Pretraining with Location-aware Captioners, Wan etal. 2024